Doerr, Martin and Apostolis Sarris (eds) 2003. The Digital Heritage of Archaeology. CAA2002. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 30th CAA Conference, Heraklion, Crete, April 2002.

"Mixed-mode" approaches to the grouping of ceramic artefacts using S-Plus

C.C. Beardah[1], M.J. Baxter[1], I. Papageorgiou[2] and M.A. Cau[3]

[1] Faculty of Science, The Nottingham Trent University,
Clifton, Nottingham NG11 8NS, United Kingdom.
{christian.beardah, michael.baxter}@ntu.ac.uk
[2] Department of Statistics, Athens University of Economics and Business,
76 Patission Str., 10434 Athens, Greece.
ioulia@aueb.gr
[3] Department of Archaeology, The University of Sheffield, Northgate House, West Street, Sheffield S1 4ET, United Kingdom.
[Address for correspondence: ERAUB, Departament de Prehistòria, Història Antiga i Arqueologia, Facultat de Geografia i Història, c/ de Baldiri Reixac s/n 08028 Barcelona, Spain.]
angelc@trivium.gh.ub.es

Abstract. The scientific analysis of ceramics can have the aim of identifying groups of similar artefacts. Separate groupings could be assumed to indicate, for example, distinct origins of the artefacts. Much published work focuses on the analysis of data derived from either geochemical or mineralogical techniques, and the former is far more likely to be subjected to quantitative statistical analysis. Our contribution to the EC funded GEOPRO Research Network has been an investigation into a "mixed-mode" approach to the quantitative statistical analysis of data arising from both kinds of techniques. This paper provides a review of our work in this area to date.
Keywords. ceramics; petrographic thin-sections; multivariate analysis; cluster analysis; mixed-mode analysis; S-Plus.

1 Introduction

At CAA2000 we presented a paper showing how, with the aid of powerful statistical software such as S-Plus (Venables and Ripley 1999), traditional methods of exploratory multivariate analysis can be used alongside, or in combination with, a technique designed specifically for grouping ceramic artefacts by chemical composition. (See Beardah and Baxter 2001, and section 2 below.) This was followed at CAA2001 by a discussion, summarised in section 3 below, of how S-Plus can be used to address issues involved in the clustering of such artefacts on the basis of categorical data arising from the analysis of petrographic thin-sections (Beardah et al. 2002). Now, in the third paper in this series, we present possible approaches to the inclusion of both petrographic and geochemical data in a statistical analysis of artefact compositional data. Two such approaches are demonstrated using data arising from the petrographic and chemical analysis of 115 specimens of Late Roman Cooking Ware from the Balearic Islands and the eastern Iberian peninsula.

In our first approach, demonstrated in section 4, the chemistry and petrography are analysed separately, but possibly concurrently, using methods appropriate for each. The nature of the S-Plus interface makes it possible to (a) easily identify sub-groups within the data and (b), compare the results when using different methods independently. Using this methodology, we can investigate whether sub-groups identified with various methods for analysing petrographic data are also identified using exploratory methods for analysing geochemical data.

Finally, in section 5, we consider the direct analysis of a combination of both types of information, treated on an equal footing. This could be achieved in a variety of ways, some of which will be discussed. For example, prior to a statistical analysis, both the chemical and petrographic data could be re-cast in the same format (either continuous or categorical). The resulting combined dataset could then be examined using appropriate techniques. Alternatively, we discuss methods that can deal directly with data of mixed type, for example the combination of continuous chemical data and petrographic data coded to reflect the presence/absence of categories.

2 Geochemical Data

The scientific analysis of the chemical composition of ceramics naturally leads to multivariate data, represented mathematically as an n by p data matrix, that is often explored using techniques such as principal components analysis (PCA) and cluster analysis. Beardah and Baxter (2001) discuss how, with the aid of the statistical software package S-Plus, traditional multivariate methods can easily be used alongside, or in combination with, a technique designed specifically for grouping ceramics by chemical composition (Beier and Mommsen 1994). This latter technique involves grouping together artefacts whose chemical compositions are "close" with respect to a mathematical measure of dissimilarity. The measure used (modified Mahalanobis distance) takes into account uncertainty of measurement and the possibility of constant shifts in the data due, for example, to dilution of the clay or instrumental variation.

Figure 1 shows output based upon the application of Beier and Mommsen's procedure on chemical compositional data for 100 ceramic samples from Nichoria in the Peloponnese. These data form part of the Perlman-Asaro databank of Mycenaean samples.

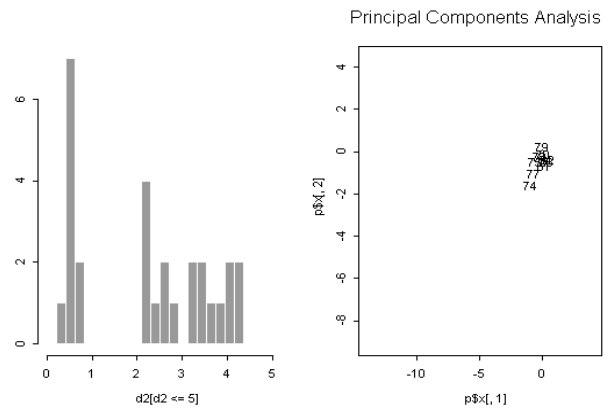


Fig. 1. *The result of growing an initial grouping of just two cases (74 and 77). The histogram of squared distances (left) reveals a pronounced "edge" at a value of less than 1 and the labelled PCA plot on the right reveals that the final grouping is quite compact.*

A preliminary scan through the data set reveals that cases 74 and 77 are very close together. Using these two objects as our initial grouping, we look for objects that are close to our current group. This process is repeated iteratively until the group stabilises. A histogram of (squared) distance values (see Figure 1) is a useful tool for identifying groups in this way and here reveals that our current group is both compact and distant from other cases. This group of ten cases is therefore classified as a genuine sub-group within these data.

3 Mineralogical Data

Beardah et al. (2002) and Cau et al. (2002) explore issues involved in clustering artefacts on the basis of thin-section data. A typical thin-section is shown in Figure 2.

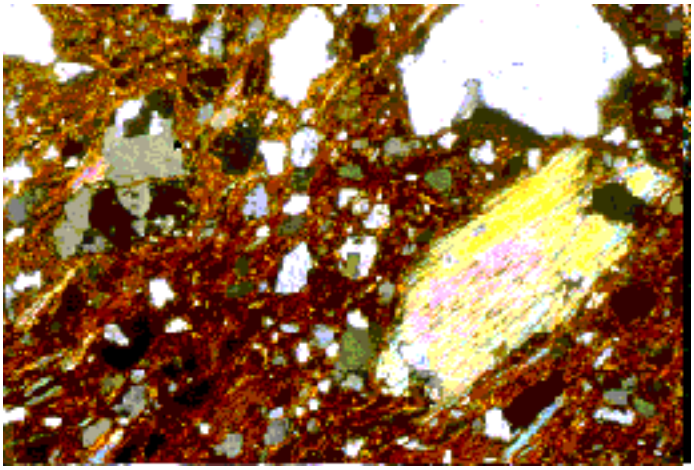
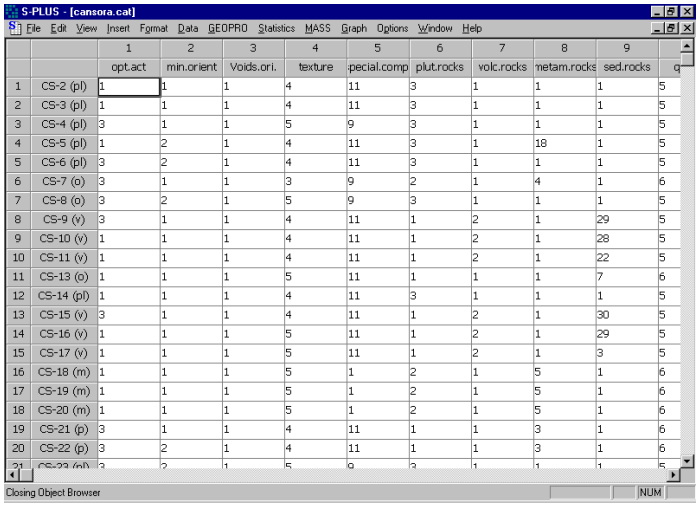


Fig. 2. *A typical thin-section.*

The first challenge we face is how to describe thin-sections in a manner that is amenable to statistical analysis. One approach is to represent each thin-section in terms of qualitative or categorical variables chosen by the analyst (see Figure 3).

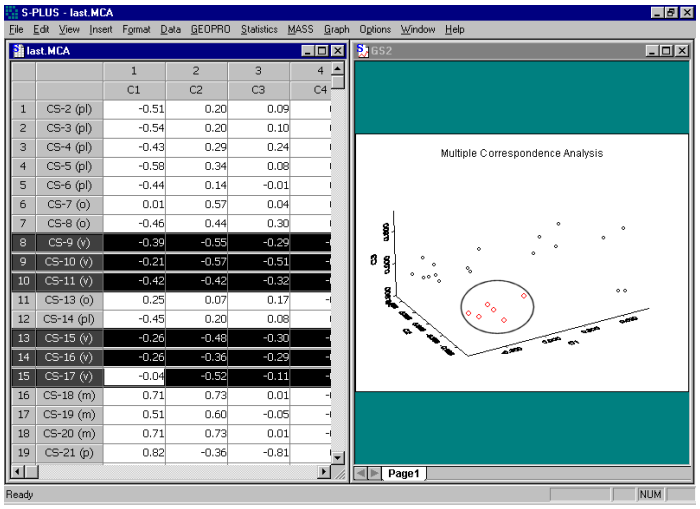
After description using such a system (see Cau et al. for an example), a collection of n thin-sections can be represented by a table with n rows and q columns. Each row contains the description of a single thin-section; each column represents a variable; and each cell contains a number indicating the category. In order to apply statistical methods, one approach is to convert the table of categorical data into a binary data matrix consisting of 0/1 entries (see Cau et al. for details).


S-PLUS - [cansora.cat]

		opt.act	min.orient	Voids.ori.	texture	pecial.comp	plut.rocks	volc.rocks	netam.rocks	sed.rocks
1	CS-2 (pl)	1	1	1	4	11	3	1	1	1
2	CS-3 (pl)	1	1	1	4	11	3	1	1	1
3	CS-4 (pl)	3	1	1	5	9	3	1	1	1
4	CS-5 (pl)	1	2	1	4	11	3	1	18	1
5	CS-6 (pl)	3	2	1	4	11	3	1	1	1
6	CS-7 (o)	3	1	1	3	9	2	1	4	1
7	CS-8 (o)	3	2	1	5	9	3	1	1	1
8	CS-9 (v)	3	1	1	4	11	1	2	1	29
9	CS-10 (v)	1	1	1	4	11	1	2	1	28
10	CS-11 (v)	1	1	1	4	11	1	2	1	22
11	CS-13 (o)	1	1	1	5	11	1	1	1	7
12	CS-14 (pl)	1	1	1	4	11	3	1	1	1
13	CS-15 (v)	3	1	1	4	11	1	2	1	30
14	CS-16 (v)	1	1	1	5	11	1	2	1	29
15	CS-17 (v)	1	1	1	5	11	1	2	1	3
16	CS-18 (m)	1	1	1	5	1	2	1	5	1
17	CS-19 (m)	1	1	1	5	1	2	1	5	1
18	CS-20 (m)	1	1	1	5	1	2	1	5	1
19	CS-21 (p)	3	1	1	4	11	1	1	3	1
20	CS-22 (p)	3	2	1	4	11	1	1	3	1

Fig. 3. *A typical categorical coding.*

Given such a binary coding, a variety of analytical options are open. These involve choosing how to measure similarity between cases and how, subsequently, to group these using clustering or scaling techniques. Options include multiple correspondence analysis (MCA) and metric and non-metric scaling methods including Sammon mapping and isotonic multidimensional scaling. All of these methods result in graphical output similar in nature to that of a PCA.



		C1	C2	C3	C4
1	CS-2 (pl)	-0.51	0.20	0.09	
2	CS-3 (pl)	-0.54	0.20	0.10	
3	CS-4 (pl)	-0.43	0.29	0.24	
4	CS-5 (pl)	-0.58	0.34	0.08	
5	CS-6 (pl)	-0.44	0.14	-0.01	
6	CS-7 (o)	0.01	0.57	0.04	
7	CS-8 (o)	-0.46	0.44	0.30	
8	CS-9 (v)	-0.39	-0.55	-0.29	
9	CS-10 (v)	-0.21	-0.57	-0.51	
10	CS-11 (v)	-0.42	-0.42	-0.32	
11	CS-13 (o)	0.25	0.07	0.17	
12	CS-14 (pl)	-0.45	0.20	0.08	
13	CS-15 (v)	-0.26	-0.48	-0.30	
14	CS-16 (v)	-0.26	-0.36	-0.29	
15	CS-17 (v)	-0.04	-0.52	-0.11	
16	CS-18 (m)	0.71	0.73	0.01	
17	CS-19 (m)	0.51	0.60	-0.05	
18	CS-20 (m)	0.71	0.73	0.01	
19	CS-21 (p)	0.82	-0.36	-0.81	

Fig. 4. *The first three component scores of an MCA (Can Sora dataset).*

For example, Figure 4 shows output from the application of the MCA technique to coded mineralogical data based upon 25 samples of Late Roman Cooking Ware (LRCW) from Can Sora (Eivissa). The output suggests the existence of several distinct subgroups, one of which is highlighted (the six cases towards the bottom of the MCA plot). Figure 3 shows part of the categorical description of these data.

4 Combining Mineralogical And Geochemical Analyses

The Can Sora data set discussed above is a subset of a more extensive data set consisting of 115 samples of LRCW from the Balearic Islands and the eastern Iberian peninsula. For these data we have both geochemical and mineralogical information. The former consists of 25 concentration values and the latter has been coded as a binary matrix. Given both kinds of information, we can concurrently apply methods appropriate to each.

As a simple example, plots of the component scores based upon PCA and MCA (applied to geochemical and mineralogical information respectively) reveal that cases CS-26 and CS-27 are clear outliers in both analyses (highlighted towards the top-right of both plots in Figure 5). Re-calculating the PCA and MCA scores with these cases omitted gives clearer plots from which further subgroups may be identified.

A Procrustes statistic developed by Sibson (1978) can be used to measure the difference between the n by k component scores resulting from separate mineralogical and geochemical analyses (for k = 1, 2, …). If X and Y are two n by k matrices of scores the statistic is defined as

1 – [{tr(YTXXTY)1/2}2/{tr(YTY)tr(XTX)}]

where tr(.) is the trace and T the matrix transpose operator. Other Procrustes statistics could be used, however this one has the merit of symmetry as it does not depend on which set of scores is designated as the Y and which the X matrix. It takes values between 0 and 1, with 0 arising for identical configurations. If the statistic is close to 0 there may be no need for a combined "mixed-mode" analysis of the kind discussed later.

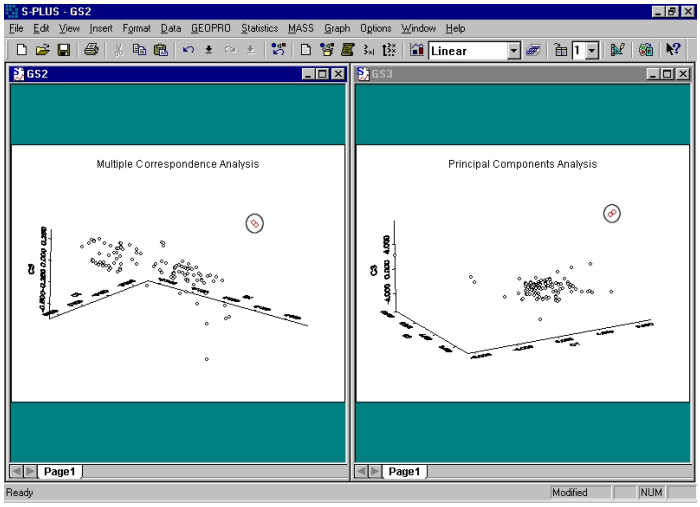


Fig. 5. *MCA and PCA of the LRCW dataset.*

This judgement may be made as part of an iterative process involving the removal of outliers from both data sets. For example, comparing PCA and MCA component scores for the LRCW data (after omitting three clear chemical and mineralogical outliers) gives the values of Sibson's coefficient shown in Table 1.

Since values near to 0 indicate similarity, here the individual analyses seem to be quite different, so we may consider combining these data in a "mixed-mode" approach.

Components, k	1	2	3
Sibson's coefficient	0.97	0.82	0.78

Table 1. *Sibson's coefficient values.*

5 "Mixed-mode" Analysis

Using separate, but possibly concurrent analyses, we can investigate whether subgroups identified on the basis of mineralogical information are also identified on the basis of geochemical information. Alternatively, some methods can deal directly with data of mixed type. To make use of such methods, we need to measure dissimilarity between objects of mixed type (Kaufman and Rousseeuw 1990). The dissimilarity coefficient used (based upon an extension of Gower's (1971) coefficient) is:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \in [0,1]$$

where dij(f) is the contribution of variable f to d(i,j) and δij(f) is the weighting of variable f and depends on the variable type. It is assumed that there is no missing data.

For our purposes, we specialize to the case where variables are binary or continuous. For continuous data

$$d_{ij}^{(f)} = \left| x_{if} - x_{jf} \right| / r_f$$

where rf is the range of variable f, so that the contribution of the variable is between 0 (identical) and 1 (most different).

Binary variables may be treated symmetrically or asymmetrically. In either case we can define dij(f) to be 0 if xif = xjf and 1 otherwise. The weights δij(f)=1 unless a variable is asymmetric binary and xif = xjf = 0, in which case it is equal to 0. Whether 0-0 matches should contribute to the dissimilarity (symmetric binary) or not (asymmetric binary) is not a purely statistical issue.

For most of the mineralogical variables used one category was "absent", and we removed the associated binary variable from the analysis, choosing not to regard mutual absence of the variable as evidence of similarity. Whether a symmetric or asymmetric treatment of other binary data was used did not make much difference to our analyses. In general this may not be the case, and our current preference is for an asymmetric treatment.

It is possible to both simplify and extend the definition of d(i,j) for our combination of binary and continuous data by writing it in the form

$$d(i,j) = \frac{B + \lambda C}{W + \lambda c} \in [0,1]$$

where b is the number of binary variables and W≤b is the number of these with non-zero weight; c is the number of continuous variables;

$$B = \sum_{f=1}^{b} \delta_{ij}^{(f)} d_{ij}^{(f)}$$

and

$$C = \sum_{h=1}^{c} d_{ij}^{(h)}$$

are the contributions to the numerator of the binary and continuous variables. The generalization arises through λ which is a weighting factor that has the value 1 in the original definition.

The S-Plus function daisy can be used to calculate a dissimilarity matrix based upon the extension of Gower's coefficient, for mixed-mode data, e.g.

"P + C" = [P | C].

Here P is a binary (or asymmetric binary) matrix containing coded mineralogical data and C is a matrix containing chemical composition data.

The output from daisy (a dissimilarity matrix) can be used as input to various clustering routines, for example Classical Metric Multi-Dimensional Scaling (CMD) or Cluster Analysis. The former technique again results in an n by k matrix of component scores that can be visualised easily for small k. In addition, using Sibson's coefficient, the output could be compared to scores arising from appropriate analyses of purely chemical or purely mineralogical data.

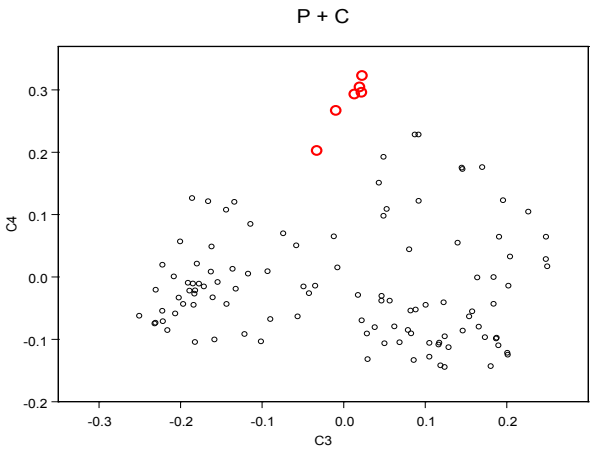


Fig 6. *A "mixed-mode" analysis of the LRCW data.*

Figure 6 shows CMD output based upon a "mixed-mode" analysis of the full LRCW dataset. The highlighted group of points towards the top of the plot corresponds to a subgroup also identified, on the basis of a "subjective" analysis of thin-sections, by Cau et al. Without prior removal of outliers, this subgroup is not so readily identified in plots based upon either chemistry or mineralogy alone.

It has been suggested that the binary data may tend to dominate analyses of this type. To date we have little evidence for this in our work. However, in order to investigate, and possibly to overcome this potential problem, we have tried giving the chemical data greater weighting by analysing P+C, P+2C, … where, for example

"P + 2C" = [P | C | C].

This is equivalent to using weights λ = 1, 2, … in the coefficient d(i,j). Again, the scores arising from these "weighted mixed-mode" analyses could be compared to those resulting from appropriate analyses of P or C alone using Sibson's coefficient.

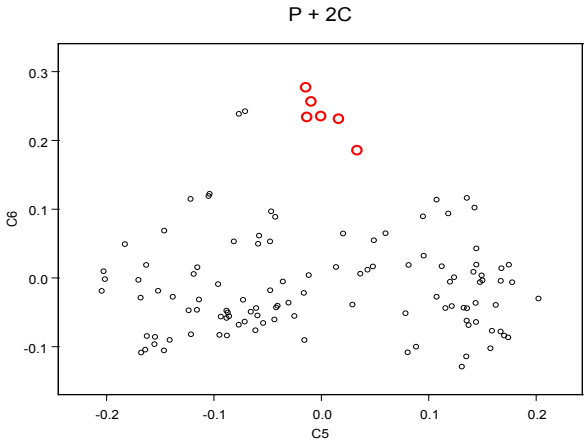


Fig. 7. *A weighted "mixed-mode" analysis of the LRCW data.*

Figure 7 shows CMD output based upon an analysis of P+2C for the full LRCW dataset. The highlighted group of points corresponds to the same subgroup identified earlier and is more readily separated in this weighted mixed-mode analysis (compare with Figure 6). Furthermore, the small outlying group consisting of cases CS-26 and CS-27 is again easily identified (to the left of the highlighted group). In Figure 8 we show CMD output after these groups have been removed. Here the highlighted group of points (towards the top-right of the plot) corresponds to a subgroup consisting of several sedimentary fabrics identified on the basis of thin-section analysis in Cau et al.

This close agreement between the "subjective" analysis of the thin-sections and the statistical methodology is only revealed by weighted mixed-mode analysis, and not by statistical analysis of either chemistry or mineralogical information alone.

6 Summary and conclusions

We have discussed various methods for the analysis of geochemical and mineralogical data, and the mixture of the two. All these methods have been implemented in the S-Plus package by us, or have been made freely available by other authors, for example as part of the MASS library (Venables and Ripley 1999). Our final collection of routines will be made freely available, via the Internet, to the archaeometric community.

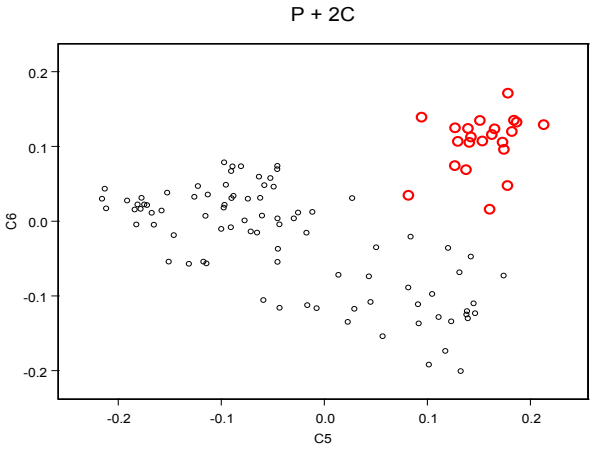


Fig. 8. *A weighted "mixed-mode" analysis of the LRCW data after removal of the groups identified on the basis of Figure 7.*

In looking at different views of the data generated by these methods, many questions can be asked. Are there clear mineralogical and chemical groups? If so, are these the same? Do any chemical groups that emerge subdivide mineralogical groups, or cross-cut them? Does a mixed-mode analysis suggest groups not apparent from single-mode analyses? For large data sets when clear groups are identified their removal from the data set may be justified, so that less obvious structure can be investigated. Essentially what is proposed here is a highly iterative approach to data exploration, the merits of which can only be determined in practice.

Graphically, one is limited (practically) to looking at two or three-dimensional configurations, but Sibson's coefficient (or something similar) can be used to effect comparisons in any number of dimensions.

Acknowledgements

This work forms part of the GEOPRO Research Network funded by the DGXII of the European Commission, under the TMR Network Programme (Contract Number ERBFMRX-CT98-0165).

References

BEARDAH, C.C. and BAXTER, M.J., 2001. Grouping ceramic compositional data: an S-plus implementation. In Stancic, Z. and Veljanovski, T. (eds.), Computing Archaeology for Understanding the Past CAA2000, Proceedings of the 28th Conference, Ljubljana, April 2000, Oxford: Archaeopress (BAR International Series 931).

BEARDAH, C.C., BAXTER, M.J., PAPAGEORGIOU, I. and CAU, M.A., 2002. Approaches to petrographic data analysis using S-Plus. In Burenhult, G. and Arvidsson, J. (eds.), Archaeological Informatics: Pushing the Envelope CAA2001, Proceedings of the 29th Conference, Gotland, April 2001, Oxford: Archaeopress (BAR International Series 1016).

BEIER, T. and MOMMSEN, H., 1994. Modified Mahalanobis filters for grouping pottery by chemical composition, Archaeometry, 36, 287-306.

CAU M.A., DAY, P.M., BAXTER, M.J., PAPAGEORGIOU, I., ILIOPOULOS, I. and MONTANA, G., 2002. Exploring automatic grouping procedures in ceramic petrology. (Submitted for publication.)

GOWER, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics, 27, 857-871.

KAUFMAN, L. and ROUSSEEUW, P.J., 1990. Finding Groups in Data. New York: John Wiley.

SIBSON, R., 1978. Studies in the robustness of multi-dimensional scaling: Procrustes statistics. Journal of the R.S.S. (B), 40, 234-8.

VENABLES, W.N. and RIPLEY, B.D., 2000. S Programming. N
ew York: Springer-Verlag.